学校编码：10384　　分类号_____密级_____
学　　号：23220111153249　　UDC _____

厦门大学

硕 士 学 位 论 文

基于方差和中值调整的 RNA-Seq 数据标准化方法及其评估

A Normalization Method Based on Variance and Median Adjustment for RNA-Seq Data and its Evaluation

王英

指导教师姓名：吉国力　教　　授
　　吴小惠　助理教授
专　业　名　称：系　统　工　程

论文提交日期：2014　年　4　月
论文答辩时间：2014　年　　月
学位授予日期：2014　年　　月

答辩委员会主席：
评　　阅　　人：

2014 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“ √ ”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

随着新一代高通量测序技术的飞快发展，RNA 测序技术（RNA-Seq）已经被广泛应用于各种生物的转录组分析中。由于不同的测序文库是由不同测序泳道产生的且测序深度有所不同，导致不同文库不能直接进行比较分析。因此，需要对测序文库序列数据进行标准化处理来调整不同测序泳道的总序列数，消除实验过程中测序技术上的误差，使能够更准确地分析基因的表达差异。

本文提出了最小方差中值标准化方法，基于方差和中值调整对 RNA-Seq 数据进行标准化，既考虑了测序文库中所有基因的表达水平对全局表达量的影响，又考虑到每个单一的基因表达量影响。本文利用该方法对拟南芥多聚腺苷化[ploy(A)]位点和基因数据集进行分析，基于几何平均法计算几何平均方差和基于加权截尾法计算均值方差，对每个样本综合两个方差得到一个最优方差，最后对方差调整后的整个测序文库的所有样本进行中值调整，从而实现数据的标准化。

本文最后基于不同样本间的数据分布、均方误差和 K-S(Kolmogorov-Smirnov)统计两个经验统计量以及差异表达分析等方式对最小方差中值标准化方法进行评估，并与已有的标准化方法 DESeq(Differential Expression Sequence)和 TMM(Trimmed Mean of M Values)进行综合比较。实验结果表明，最小方差中值标准化方法能用于有效处理高通量 RNA-Seq 数据，实现不同条件下的测序数据的标准化，使标准化后的各个测序文库序列数样本具有相同数据分布，并能将所有测序样本调整到同一水平，缩小了测序文库中基因和 poly(A)位点在不同测序样本下的总体表达差异。

关键词：RNA 测序；标准化方法；评估

Abstract

With the rapid development of the next generation sequencing, the RNA sequencing(RNA-Seq) is widely used in the analysis of transcriptomes of various organisms. As the different sequencing libraries are gained by the different sequencing lanes and the sequencing depths are different, these different libraries can not be directly compared. Therefore, these sequencing libraries must be normalized to adjust the total number of the different sequencing lanes to eliminate the errors of sequencing technology in the experimental process and to enable more accurate analysis of differentially expressed genes.

The paper proposes the minimum variance and median normalization method, which is based on variance and median adjustment to normalize the RNA-Seq datasets. This method not only considers the global expression level of all genes in the overall library, but also considers the impact of each individual gene's expression. This method is used to analyze the Arabidopsis polyadenylation[ploy(A)] and gene datasets. First, the geometric mean variance based on geometric mean method is calculated, and the mean variance is calculated based on the weighted trimmed. Then these two variances are synthesized for each sample to obtain an optimal variance. Finally median adjustment is performed for all the samples from the sequencing libraries after variance adjustment to achieve the normalization of datasets.

In this paper, the minimum variance and median normalization method is evaluated based on the data distribution of different samples, the empirical statistical metrics[mean square error(MSE) and Kolmogorov-Smirnov(K-S) statistic], differential expression analysis and so on. The minimum variance and median normalization method is also compared synthetically with the two exiting normalization methods, DESeq(Differential Expression Sequence) and TMM(Trimmed Mean of M Values). The experimental results show that the minimum variance and median normalization method can effectively normalize the RNA

high-throughput datasets under different conditions, make each sample of the normalized sequencing library have the same data distribution, adjust the all sequencing samples to the same level, and reduce the overall expression difference of genes and poly(A) sites in different samples from sequencing library.

Keywords: RNA Sequencing; Normalization Method; Evaluate

目 录

Content

第一章 绪论

1.1 研究背景及意义

高等真核生物的细胞生命活动过程中，基因 DNA 分子中储存的特定遗传物质信息，可以经过转录和加工修饰过程，最终由成熟的信使 RNA（mRNA）翻译成具有特定结构和功能的活性蛋白质大分子物质[1, 2]。多聚腺苷化（Polyadenylation）是指真核细胞基因表达过程中，转录得到的产物信使 RNA 前体（Pre-mRNA）经过加工处理形成成熟的信使 RNA 过程中的 3’末端的多聚腺苷化[3]。具体是指初始转录物的 3’端的某一个特定的位置[poly(A)位点]经过剪切（Cleavage）操作被去掉一小段序列，从而得到一个新的 3’末端，同时在这个新的末端处添加一个由 50~250 个腺嘌呤组成的多聚 A 尾巴[AAAAAA……，poly(A)尾巴]。多聚腺苷化得到的产物成熟的信使 RNA 在辅助因子的作用下，经过细胞质进入核糖体，最终翻译成生物蛋白质分子。成熟的 mRNA 的多聚 A 尾巴可以保护其免受核酸外切酶的攻击，决定信使 RNA 的整个生命活动周期，并促进信使 RNA 从细胞核输出转移、促进信使 RNA 翻译成蛋白质以及防止信使 RNA 降解[4]。Poly(A)位点不仅决定了成熟的信使 RNA 的外显子和调控元件，而且标志着基因编码蛋白质的终止位置。正确识别 poly(A)位点，有助于生物研究者对高等真核生物进行基因识别、基因组注释分析以及探索基因表达过程中的调控机制等。

研究表明，绝大多数高等真核生物基因中含有大于一个的 poly(A)位点，基因表达过程中对这些位点的差异选择和不同的剪接方式，使信使 RNA 具有多样性，从而最终可能改变其翻译产生的活性蛋白质分子的结构和功能[5, 6]。基因表达过程中发生的对多个 poly(A)位点的差异选择和使用现象，称为选择性多聚腺苷化（Alternative Polyadenylation，APA）[7]。选择性多聚腺苷化会影响真核生物基因的表达，影响生物个体的生长、发育、代谢等生命活动过程，还可能引起生物个体发生病变等[8]。

近年来，随着人类基因组计划等多个跨国合作项目的深入开展，人类积累了

大量的生物信息数据，特别是关于细胞核酸（脱氧核糖核酸和核糖核酸）和蛋白质等生物大分子的碱基序列、分子结构及其功能等方面的数据，对生物个体的基因组结构和功能有了一个比较系统的认识和理解。如何分析和处理这些大规模的生物数据信息，如何快速而又准确的挖掘这些测序数据蕴含的基因遗传物质信息，对揭示生物的起源与进化和探索生命的奥秘都有着非常重要的意义。

测序技术是一种能够破译隐藏在生物基因组 DNA 碱基序列中的遗传物质，为科研工作者研究生物个体基因组的遗传物质信息的结构和功能、全面解释和分析生物的复杂性和多样性直接提供生物信息数据的技术[9]。新一代高通量测序技术在测序成本以及测序时间等方面的突破，使这一技术成为了一般生物研究实验室都可以使用和拥有的技术，成为现代分子生物学家进行基因转录组学、表现基因组学以及基因组学等相关学科研究的主要工具[10, 11]。

利用新一代高通量测序技术对各种生物类型的转录本进行定量的深度检测，这就是所谓的 RNA 测序（RNA Sequencing，RNA-Seq）技术。RNA-Seq 技术[12-14]是目前用于转录组和基因表达研究的一种新的生物信息学分析手段，由于其独特的优势，已经被广泛的应用于拟南芥、水稻以及人类和其他生物的相关研究中。利用 RNA-Seq 进行生物的转录组研究，有助于人类发现新的基因、揭示基因表达过程以及研究结构变异等。

新一代高通量测序技术可以产生大规模的生物测序数据，对这些测序数据信息进行合理的分析和挖掘，消除测序技术偏差对实验样本数据带来的误差在整个数据统计分析中具有重要的意义[15]。本文主要围绕对 RNA-Seq 产生的拟南芥模式植物 poly(A)位点和基因两个测序数据集进行标准化和差异表达分析，以及如何通过经验统计量等分析标准化后各个样本数据的质量，从而对标准化方法的可行性和有效性进行综合评估这两个内容开展研究工作的。

1.2 研究现状及内容

测序文库序列数只有经过有效的数据标准化处理，才能提取出其中真正具有生物意义的表达观测值，进而对其进行精确估计和差异表达分析，从而为解决后续的生物问题提供定量的分析依据[16]。RNA-Seq 数据标准化的目的是为了消除